Editorial

Identifying and Assessing the Null Hypothesis

Most research studies develop an implicit or explicit model, derive one or more hypotheses, and then use statistical tests to support or reject the hypotheses. In many instances, the investigator tests the relatively simple hypothesis that a variable has no effect and then focuses on the statistically significant results. This is not necessarily the correct approach, at least from the perspective of editors and a readership interested in policy-relevant findings.

Several issues are worth considering before undertaking the analysis and interpreting the data. First, what really is the null hypothesis? I believe that most people are, to some extent, implicit Bayesians. That is to say, even if we are open-minded, we have expectations concerning likely results based on theory, prior research, experience, hunches, and prejudice. Although statistical software is now so easy to use, few would knowingly exert even that minimal effort, truly believing that none of the variables had any effect.

Except for the first study asking an entirely new question, there is a body of research that is relevant and that should be considered and carefully reviewed. Well-executed studies cite the evidence for *and* against the proposed relationship. In some instances, the prior evidence is consistently in one direction. If so, this would argue for a one-tailed test. In other instances, the prior evidence is less conclusive and the null hypothesis may truly be zero.

Is the finding of "a statistically significant effect" *per se*, always interesting? Probably not, unless the finding is clinically relevant or relevant to policy. The test whether an estimated coefficient is very unlikely to be zero—the classic interpretation of a statistically significant finding—is based on the estimated values of the coefficient and its variance. The variance, in turn, depends on the sample size. If the sample size is enormous, the standard error around the coefficient is very small, so nearly all coefficients may be statistically significant. There may be an effect, but is it large enough to be of interest?

We may be interested only if an intervention results in at least a 10 percent improvement in outcomes. The 10 percent figure may be derived from a sense of the costs—economic, political, or social—required by the

intervention. Or it may reflect the perception that we already have many ideas awaiting implementation that yield more than a 10 percent benefit. Regardless of how we obtain our threshold level, we can then adapt the analysis to test whether the effect is significantly greater than 10 percent, rather than zero.

The converse situation is also worth discussing. Are non-significant findings always uninteresting? Again the answer is, not always. Non-significant results may arise from many situations. The underlying methods may be flawed, the data may be collected with error, or the hypothesis may be trivial. On the other hand, if a study is well conceived and executed, non-significant findings may be due to insufficient statistical power. That is, the coefficient may suggest an important effect (from a policy perspective), but the small sample size results in confidence intervals too wide to reject the null hypothesis.

If all studies could be prospective and all budgets were unlimited, good investigators would always have sufficient sample size. However, samples are often limited by budget, the realities of a "natural experiment," or the unexpected loss of data in even a well-designed study. Is there hope for investigators in such unlucky situations? Perhaps, especially if the result can be placed in context. Suppose the question is important and no other research is available to inform policymakers. A well-done study indicating large, although not statistically significant, effects may be better than no information at all. If there are five to ten other studies, all with insignificant results, one more similar study may not add much, but if combined in formal evidence-based review, the sample size problem may be overcome. The literature on systematic reviews is growing rapidly, although it is still largely based in the clinical realm rather than in health services research more generally (Mulrow and Cook 1998). Non-significant findings are also valuable when they can *rule out* the presence of an important effect. That is, if the power to detect small differences is high, then a study demonstrating that the effect is clearly *less* than some policy-relevant value can be very valuable.

The foregoing discussion focuses on the effect of a key independent variable on the dependent variable and reflects the classic situation in which all other factors are appropriately controlled, either by randomization or appropriate statistical controls. Randomization is best, but in most situations of interest to a health services researcher, it is not feasible. Devising an appropriate statistical model is both conceptually challenging and difficult to undertake. Thus, assessing the null hypothesis of a set of empirical findings is quite a challenge.

Several articles in the current issue illustrate some of the problems and

opportunities in empirical research. For instance, in the context of major efforts by hospitals and other firms to change their work processes to become more efficient, Walston, Burns, and Kimberly (2000) ask, "Does Reengineering Really Work?" Reengineering was a major management focus (some might say fad) in the early and mid-1990s. It was "the fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in critical, contemporary measures of performance, such as cost, quality, service, and speed" (Hammer and Champy 1993). The investigators surveyed all 2,306 short-term general hospitals in urban areas with over 100 beds, with usable responses from 29.4 percent of the CEOs. They assessed the impact of restructuring and reengineering on the relative cost of the hospital within its market area.

The authors' findings indicate statistically significant *increases* in relative cost per patient day between 1993 and 1996 for hospitals implementing reengineering by 1994. These adverse consequences were reduced somewhat if the reengineering was combined with codification of the efforts and incorporated steering committees or dedicated project teams. Such offsets at best reduced the reengineering effect to a neutral impact on relative cost change.

Given the few objective studies of how well reengineering works, these are important results. They raise the bar of credibility for those encouraging major changes in hospital organization, suggesting that such efforts may actually be counterproductive. As with most complicated research studies, however, this one also raises a host of questions for further examination. One is the choice of the dependent variable: reengineering may have focused on issues other than cost, such as quality or patient satisfaction. Even in the cost realm, improved processes arising from reengineering might have focused largely on length of stay, for example by reducing the number of "handoffs" involved. If so, cost per case could be reduced but cost per day would increase because most expenses occur early in the stay.

A more complicated issue arises because of the potential for selection. The authors address the issue of whether *respondents* may represent a biased sample of the targeted hospitals. Selection is an even greater concern, however, with respect to which hospitals chose to reengineer. If we think about what might lead hospitals to implement a massive organizational change, it *could* be unrelated to relative cost trends, but it is far more plausible that it would occur among hospitals perceiving their situation as getting worse, than among those whose relative situation was getting better. If so, reengineering could have slowed the rate of increase but not enough to keep costs from still rising.

In this instance, the null hypothesis "does relative cost rise or fall?" should be stated in a more complicated way: "does relative cost rise (or fall) *more than would have been the case had reengineering not been tried*?" In the classic controlled trial, we can rely on randomization to assure—if samples are large enough and trials repeated often enough–that the experimental and control groups are identical, and courses (of illness, or relative cost) will be the same. In observational studies, we have to be more careful to control for such selection problems. This, however, requires a far more complicated research design.

In "The Impact of Utilization Management on Readmissions Among Patients with Cardiovascular Disease," Lessler and Wickizer (2000) examine a quite different question: whether the approval of fewer hospital days than requested has an impact on readmission. The authors have data on over 3,000 patients, over 80 percent of whom had the total number of days requested approved by the Utilization Management (UM) program, roughly 11–12 percent were approved for one day less than requested, and 7–8 percent had stays requested for them reduced by two or more days. The dependent variable was whether a readmission occurred within 60 days of the index admission.

No association was found between restricted length of stay and 60-day readmission rates for medical admissions. However, a small (non-significant) effect was found for procedurally focused admissions reduced by one day, and a statistically significant effect for patients whose stays were reduced by two or more days. Using a Cox proportional hazards model, the latter effect indicated a 2.6 times greater likelihood of readmission within 60 days. The regression included indicators of whether the initial admission was for cardiac catheterization, and this variable was significantly related to readmission.

Again, we have the problem of non-random assignment. That is, we do not know the likely clinical course for individuals had they been granted as many days as their doctor requested. One can always argue that non-random assignment may be a problem, but in this instance there may be other reasons to be suspicious. Even with over 3,000 cases, crucial parts of the analysis are based on relatively few patients. Among the procedural admissions, there were 805 with no reduction in stay and these had a 12.4 percent readmission rate. The results of concern arise from 82 patients with 2+ days denied, who had a readmission rate of 14.6 percent. Suppose that there was one fewer readmission—11 instead of 12; then the readmission rate would be 13.4 percent. Two fewer readmissions would bring the rate below that of the "control" group.

Aside from potentially unstable results arising from small numbers, there are other questions worth exploring. Medical and procedural patients had almost identical proportions of reductions in stay, yet no effects are observed for the medical admissions even though these are likely to be more clinically unstable. Of the procedural admissions, 30 percent were for catheterization, which is often followed by bypass surgery or angioplasty, and 17 percent were for bypass surgery, which is sometimes followed by catheterization to determine graft patency. In neither case is the readmission an unequivocal indication of failure, and in many instances the second admission is "avoided" by bundling the two together and having a somewhat longer stay. In some parts of the country, capacity constraints in hospitals lead clinicians to postpone discharge in order to maintain a position in the queue. In other areas, excess capacity allows discharge and an easily scheduled readmission. A substantial fraction of the "excess" readmissions occur in the first few days. It would be useful to see whether these were for "follow-up procedures" and also to see whether the rate of multiple procedures on the index admission was comparable for cases with and without reductions in requested stay. For policy purposes, we need to determine whether readmission rates reflect quality of care or scheduling patterns.

On the other hand, other important quality of care issues may exist that should be explored, but much larger data sets may be needed. For example, one possibility is that early discharge pressured by insurer denials leads to deterioration in patients' health status. The authors are appropriately cautious in reaching such a conclusion, since we do not know the reasons for the readmissions, let alone whether the patients were in jeopardy. An alternative possibility is that quality of care varies among clinicians and those who are less skilled ask for extended lengths of stay, perhaps based on their prior experience. Under this hypothesis, the UM program, although not through any great clinical acumen, inadvertently targets patients at greater risk because their physicians (who are less skilled) are requesting overly long stays. Whether these patients would have had fewer problems if granted longer stays is unknown, but would definitely be worth investigating.

Fisher and colleagues address quite a different problem in their article, "Associations Among Hospital Capacity, Utilization, and Mortality of U.S. Medicare Beneficiaries, Controlling for Sociodemographic Factors" (Fisher, Wennberg, Stukel, et al. 2000). They begin with the oft-observed variability in hospital use across geographic areas and ask whether sociodemographic factors might explain this, and then whether mortality is associated with differential availability and use. Their data allow admissions to be attributed

to the place of residence; therefore making selection less of a problem, but it may still be that areas with more underlying illness have built more hospital resources. Fisher et al. include some sociodemographic factors likely to be associated with underlying illness, and they go further to examine the effects for various subgroups of Medicare beneficiaries, such as African Americans and persons living in low-income areas. They also include the six-month prior hospital use rates for those who died, arguing that this subgroup was likely to be much sicker and thus more uniform across areas.

The data set they use is enormous—5.53 million Medicare beneficiaries —roughly 20 percent of the total eligible population. Thus, lack of precision is not a major problem and the estimated confidence intervals are quite small. Not surprisingly, increased availability is associated with increased utilization, even after including as many of the sociodemographic explanations as possible. The focus of the article, however, is on the observation that this increased availability is associated with *higher,* rather than lower mortality rates. This is clearly a source of potential concern, but it is important to note that even the largest effect is an odds ratio of 1.08—a slightly higher risk of death, not an 8 percentage point increase. As the authors note, they do not assess other outcomes, and in many instances, such as hip replacement, improved function comes only with some risk of near-term death. The huge sample allows the authors to avoid any concern that their research could not detect small but important improvements in mortality, as would be the case if the sample were small and confidence intervals wide. We need not agree that death rates are *higher* in well-endowed areas—if they are, the effects are small—but instead we should ask what it is that we, as individuals and a society, expect to get from the additional medical resources consumed in such areas.

In many instances it is impossible to obtain the necessary sample or the appropriate variables to truly measure the effects of interest. Nevertheless research moves forward by the slow accretion of ever-improving studies. Even if a specific design is limited, the investigators should always keep in mind, and share with their readers, what the null hypothesis truly is and what compromises they need to make in their study. Clearer definitions and assessments of the null hypothesis will benefit us all.

Harold S. Luft, Ph.D.
Institute for Health Policy Studies
University of California, San Francisco,
Senior Associate Editor of *Health Services Research*

REFERENCES

Fisher, E. S., J. E. Wennberg, T. A. Stukel, J. S. Skinner, S. M. Sharp, J. L. Freeman, and A. M. Gittelsohn. 2000. "Associations Among Hospital Capacity, Utilization, and Mortality of U.S. Medicare Beneficiaries, Controlling for Sociodemographic Factors." *Health Services Research* 34 (6): 1351–62.

Hammer, M., and J. Champy. 1993. *Reengineering the Corporation: A Manifesto for Business Revolution.* New York: HarperCollins, p. 32.

Lessler, D. S., and T. M. Wickizer. 2000. "The Impact of Utilization Management on Readmissions Among Patients with Cardiovascular Disease." *Health Services Research* 34 (6): 1315–30.

Mulrow, C., and D. Cook, eds. 1998. *Systematic Reviews: Synthesis of Best Evidence for Health Care Decisions.* Philadelphia, PA: American College of Physicians.

Walston, S. L., L. R. Burns, and H. R. Kimberly. 2000. "Does Reengineering Really Work?" *Health Services Research* 34 (6): 1363–88.